Introduction to Statistical Machine Translation

ESSLLI 2005

Chris Callison-Burch
Philipp Koehn

A long history

- Machine translation was one of the first applications envisioned for computers
- Warren Weaver (1949)

  "I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text."
- First demonstrated by IBM in 1954 with a basic word-for-word translation system.

Commercially Interesting

- U.S. has invested in MT for intelligence purposes
- MT is popular on the web -- it is the most used of Google's special features
- EU spends more than €1,000,000,000 on translation costs each year. (Semi-) automating that could lead to huge savings

Academically Interesting

- Machine translation requires many other NLP technologies
- Potentially: parsing, generation, word sense disambiguation, named entity recognition, transliteration, pronoun resolution, natural language understanding, and real-world knowledge

What makes MT hard?

- Word order
- Word sense
- Pronouns
- Tense
- Idioms

Various approaches

- Word-for-word translation
- Syntactic transfer
- Interlingual approaches
- Controlled language
- Example-based translation
- Statistical translation

Statistical machine translation

- Find most probable English sentence given a foreign language sentence
- Automatically align words and phrases within sentence pairs in a parallel corpus
- Probabilities are determined automatically by training a statistical model using the parallel corpus

Probabilities

- Find most probable English sentence given a foreign language sentence

$$p(e|f)$$

$$\hat{e} = \arg\max_{e} \; p(e|f)$$

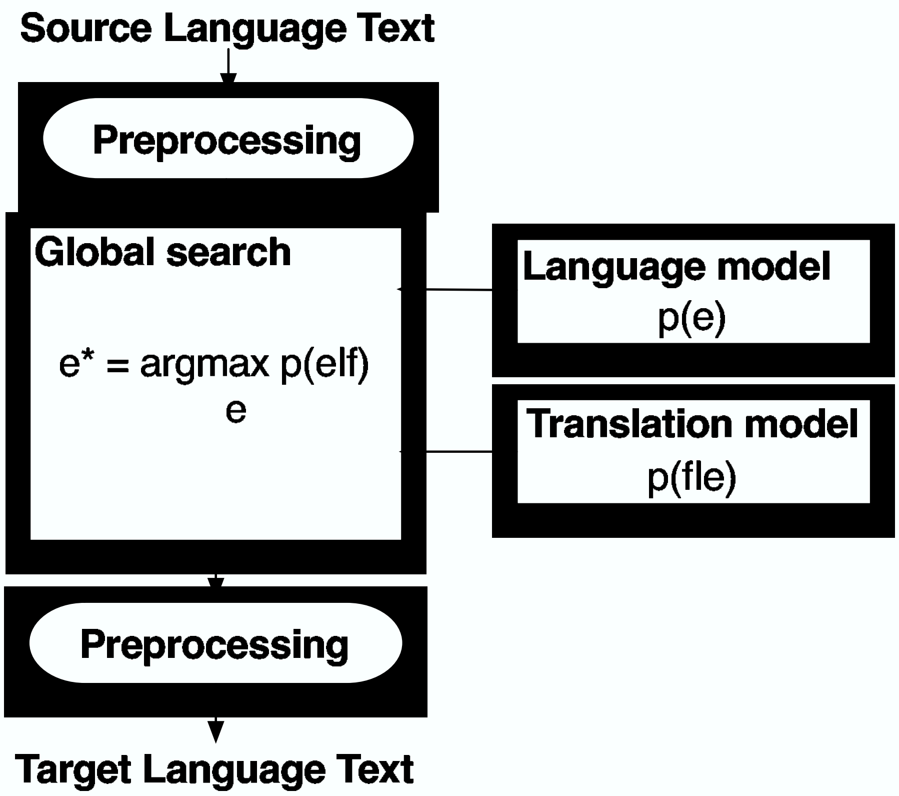
$$p(e|f) = \frac{p(e)p(f|e)}{p(f)}$$

$$\hat{e} = \arg\max_{e} \; p(e)p(f|e)$$

What the probabilities represent

- p(e) is the "Language model"
    - Assigns a higher probability to fluent / grammatical sentences
    - Estimated using monolingual corpora
- p(f|e) is the "Translation model"
    - Assigns higher probability to sentences that have corresponding meaning
    - Estimated using bilingual corpora

For people who don't like equations



Language Model

- Component that tries to ensure that words come in the right order
- Some notion of grammaticality
- Standardly calculated with a trigram language model, as in speech recognition
- Could be calculated with a statistical grammar such as a PCFG

Trigram language model

- p(I like bungee jumping off high bridges) =
  p(I | <s> <s>) *
  p(like | I <s>) *
  p(bungee | I like) *
  p(jumping | like bungee) *
  p(off | bungee jumping) *
  p(high | jumping off) *
  p(bridges | off high) *
  p(</s> | high bridges) *
  p(</s> | bridges </s>)

Calculating Language Model Probabilities

- Unigram probabilities

$$p(w_1) = \frac{count(w_1)}{total\ words\ observed}$$

Calculating Language Model Probabilities

- Bigram probabilities

$$p(w_2|w_1) = \frac{count(w_1 w_2)}{count(w_1)}$$

Calculating Language Model Probabilities

- Trigram probabilities

$$p(w_3|w_1 w_2) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)}$$

Calculating Language Model Probabilities

- Can take this to increasingly long sequences of n-grams
- As we get longer sequences it's less likely that we'll have ever observed them

Backing off

- Sparse counts are a big problem
- If we haven't observed a sequence of words then the count = 0
- Because we're multiplying the n-gram probabilities to get the probability of a sentence the whole probability = 0

Backing off

- $$\begin{aligned}&.8 * p(w_3|w_1 w_2) + \\ &.15 * p(w_3|w_2) + \\ &.049 * p(w_3) + \\ &.001\end{aligned}$$
- Avoids zero probs

Translation model

- p(f|e)... the probability of some foreign language string given a hypothesis English translation
- f = Ces gens ont grandi, vécu et oeuvré des dizaines d'années dans le domaine agricole.
- e = *Those people have grown up, lived and worked many years in a farming district.*
- e = *I like bungee jumping off high bridges.*

Translation model

- How do we assign values to p(f|e)?
- $$p(f|e) = \frac{count(f,e)}{count(e)}$$
- Impossible because sentences are novel, so we'd never have enough data to find values for all sentences.

Translation model

- Decompose the sentences into smaller chunks, like in language modeling
- $$p(f|e) = \sum_a p(a, f|e)$$
- Introduce another vairable *a* that represents alignments between the individual words in the sentence pair

Word alignment

	Those	people	have	grown	up	,	lived	and	worked	many	years	in	a	farming	district	.
Ces	■															
gens		■														
ont			■													
grandi				■	■											
,						■										
vécu							■									
et								■								
oeuvré									■							
des										■						
dizaines										■						
d'										■						
années											■					
dans												■				
le													■			
domaine															■	
agricole														■		
.																■

Alignment probabilities

- So we can calculate translation probabilities by way of these alignment probabilities

$$p(f|e) = \sum_{a} p(a, f|e)$$

- Now we need to define p(a, f | e)

$$p(a, f|e) = \prod_{j=1}^{m} t(f_j|e_i)$$

Calculating $t(f_j|e_i)$

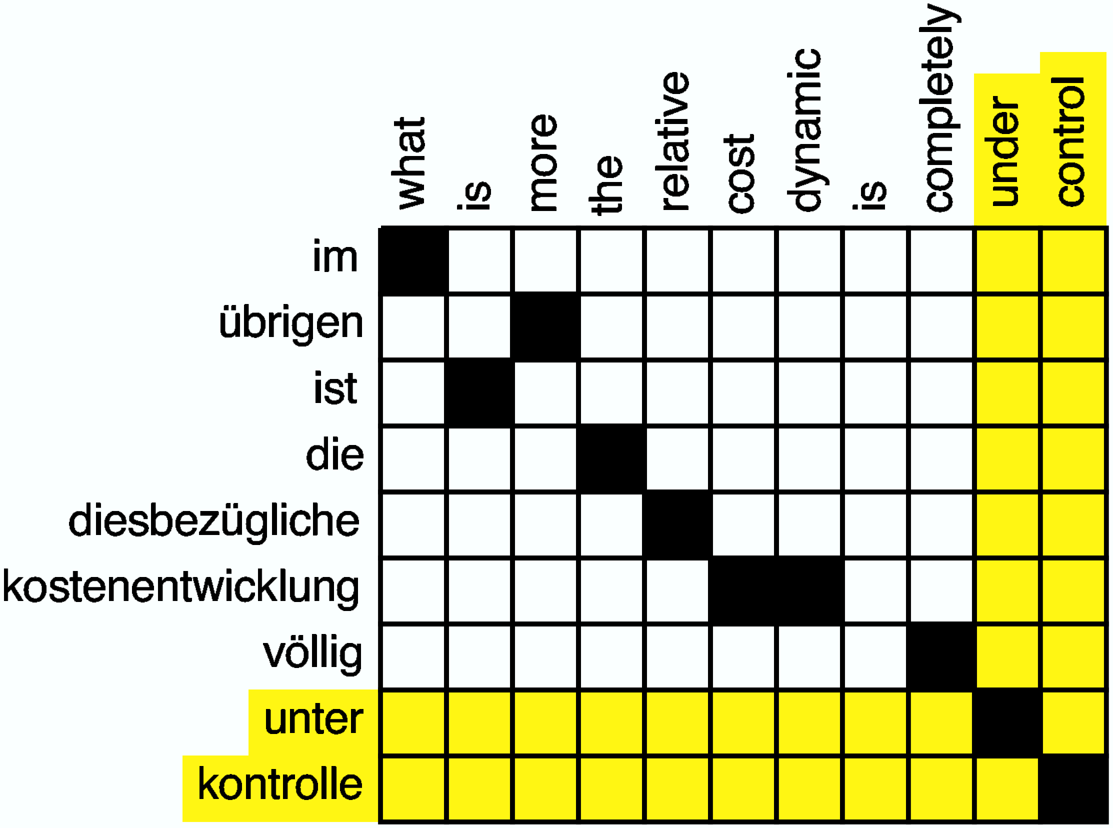
	Those	people	have	grown	up	,	lived	and	worked	many	years	in	a	farming	district	.
Ces	■															
gens		■														
ont			■													
grandi				■	■											
,						■										
vécu							■									
et								■								
oeuvré									■							
des										■						
dizaines										■						
d'										■						
années											■					
dans												■				
le													■			
domaine															■	
agricole														■		
.																■

- Counting! I told you probabilities were easy!
- $= \dfrac{count(f_j, e_i)}{count(e_i)}$
- worked... fonctionné, travaillé, marché, oeuvré
- 100 times total 13 with this f. 13%

Calculating $t(f_j|e_i)$

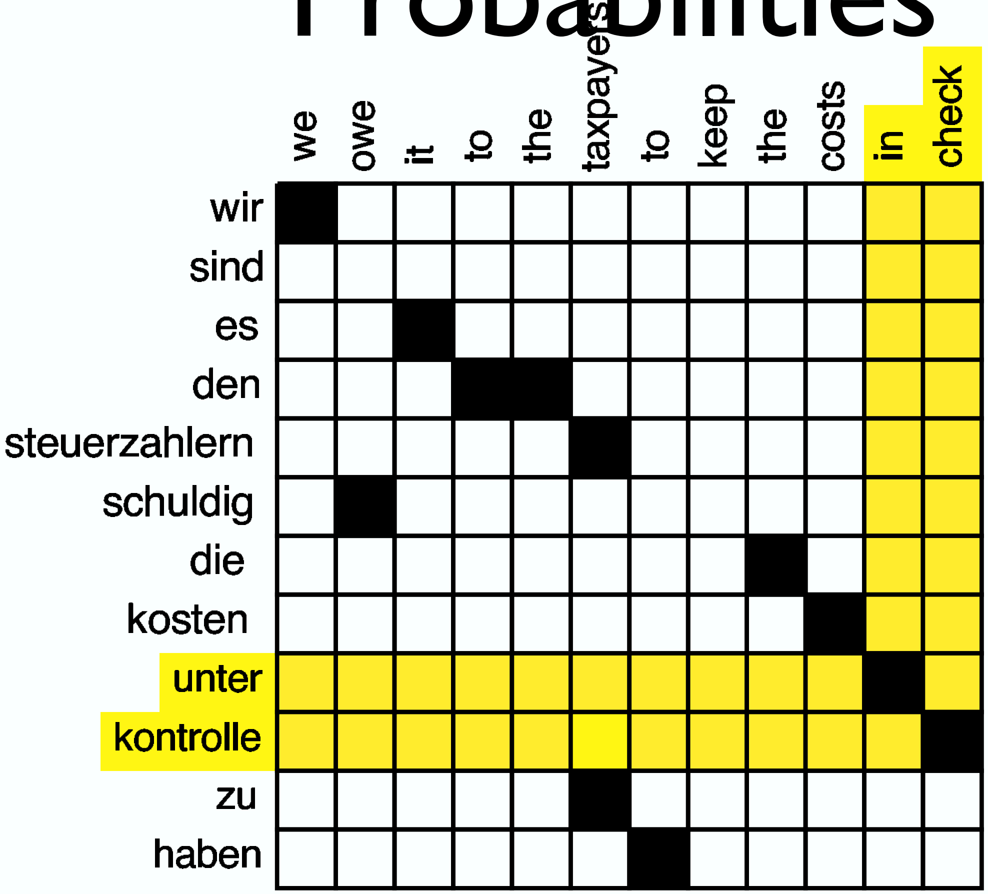
- Unfortunately we don't have word aligned data, so we can't do this directly.
- OK, so it's not quite as easy as I said.
- Philipp will talk about how to do word alignments using EM on Wednesday.

Phrase Translation Probabilities



	what	is	more	the	relative	cost	dynamic	is	completely	under	control
im	■									🟨	🟨
übrigen			■							🟨	🟨
ist		■								🟨	🟨
die				■						🟨	🟨
diesbezügliche					■					🟨	🟨
kostenentwicklung						■	■			🟨	🟨
völlig									■	🟨	🟨
unter	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	■	🟨
kontrolle	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	■

Phrase Translation Probabilities



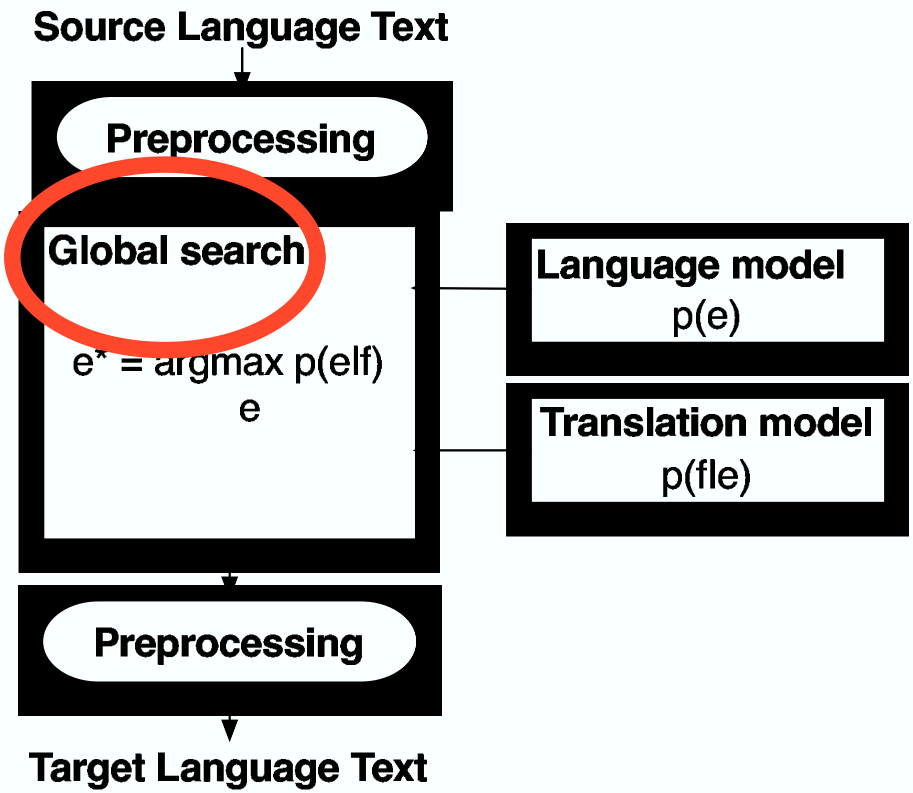
	we	owe	it	to	the	taxpayers	to	keep	the	costs	in	check
wir	■											
sind												
es			■									
den				■	■							
steuerzahlern						■						
schuldig		■										
die									■			
kosten										■		
unter											■	
kontrolle												■
zu						■						
haben							■					

Phrase Table

- Exhaustive table of source language phrases paired with their possible translations into the target language, along with probabilities

das thema	the issue	.51
	the point	.38
	the subject	.21

``Diagram Number 1"



Source Language Text

Preprocessing

Global search

e* = argmax p(elf)
e

Language model
p(e)

Translation model
p(fle)

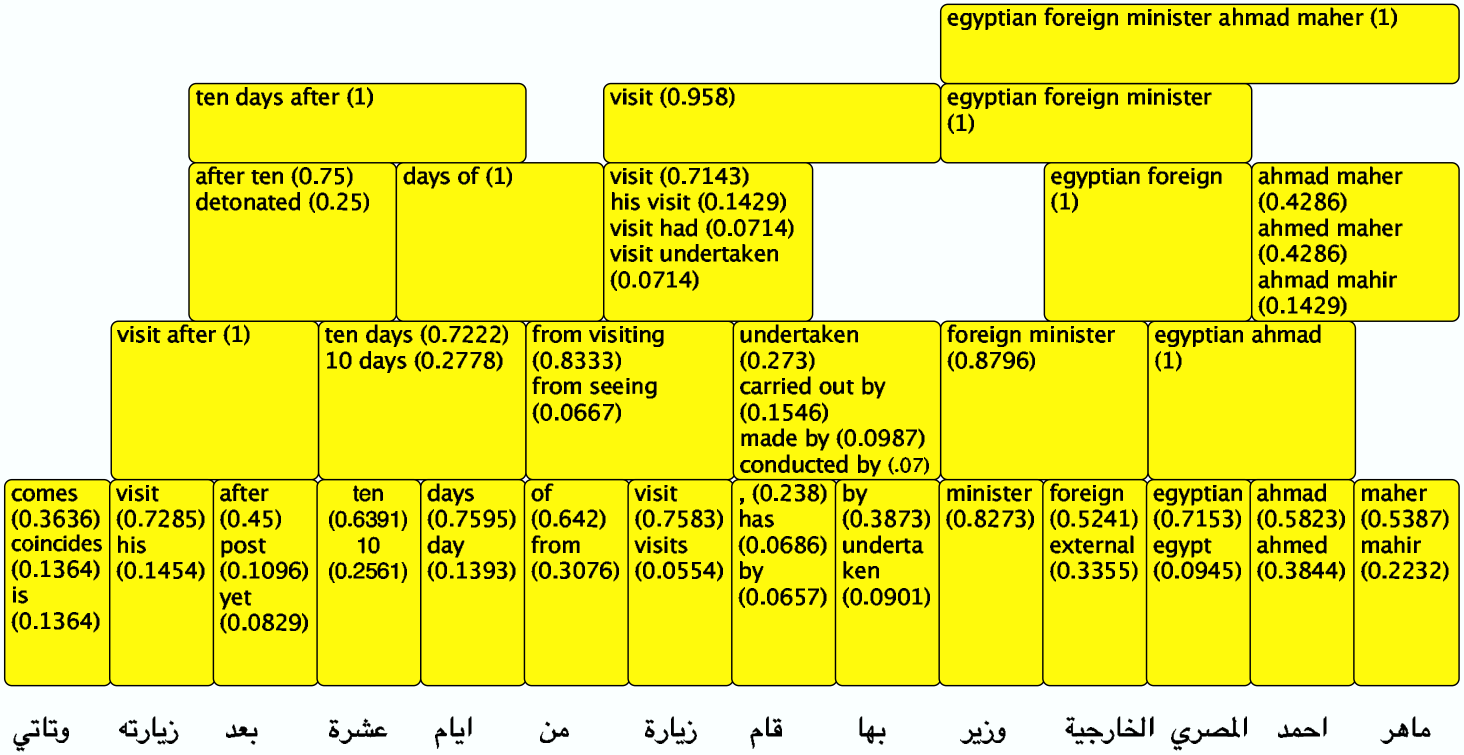
Preprocessing

Target Language Text

The Search Process AKA ``Decoding"

- Look up all translations of every source phrase, using the phrase table
- Recombine the target language phrases that maximizes the translation model probability * the language model probability
- This search over all possible combinations can get very large so we need to find ways of limiting the search space

Looking up translations of source



egyptian foreign minister ahmad maher (1)

ten days after (1)

visit (0.958)

egyptian foreign minister (1)

after ten (0.75) detonated (0.25)

days of (1)

visit (0.7143) his visit (0.1429) visit had (0.0714) visit undertaken (0.0714)

egyptian foreign (1)

ahmad maher (0.4286) ahmed maher (0.4286) ahmad mahir (0.1429)

visit after (1)

ten days (0.7222) 10 days (0.2778)

from visiting (0.8333) from seeing (0.0667)

undertaken (0.273) carried out by (0.1546) made by (0.0987) conducted by (.07)

foreign minister (0.8796)

egyptian ahmad (1)

comes (0.3636) coincides (0.1364) is (0.1364)

visit (0.7285) his (0.1454)

after (0.45) post (0.1096) yet (0.0829)

ten (0.6391) 10 (0.2561)

days (0.7595) day (0.1393)

of (0.642) from (0.3076)

visit (0.7583) visits (0.0554)

, (0.238) has (0.0686) by (0.0657)

by (0.3873) underta ken (0.0901)

minister (0.8273)

foreign (0.5241) external (0.3355)

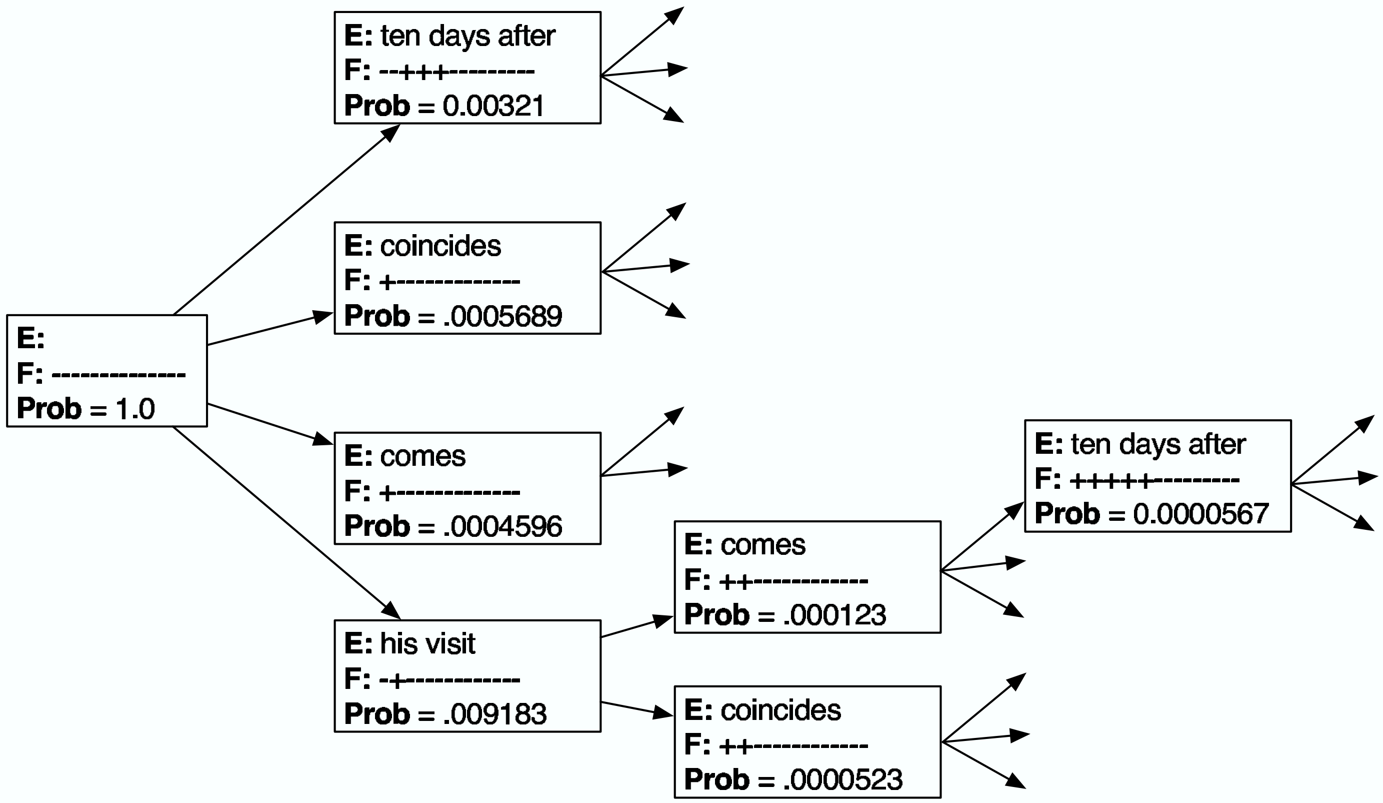
egyptian (0.7153) egypt (0.0945)

ahmad (0.5823) ahmed (0.3844)

maher (0.5387) mahir (0.2232)

وتاتي زيارته بعد عشرة ايام من زيارة قام بها وزير الخارجية المصري احمد ماهر

The Search Space


E:
F: --------------
Prob = 1.0
E: ten days after
F: --+++---------
Prob = 0.00321
E: coincides
F: +------------
Prob = .0005689
E: comes
F: +------------
Prob = .0004596
E: his visit
F: -+-----------
Prob = .009183
E: comes
F: ++-----------
Prob = .000123
E: coincides
F: ++-----------
Prob = .0000523
E: ten days after
F: +++++---------
Prob = 0.0000567